Data classification method based on network dynamics analysis and cloud model

WANG Xiao-xia, YANG Feng-bao, LIANG Ruo-fei, ZHANG Wen-hua

(*School of Information and Communications Engineering, North University of China, Taiyuan 030051, China*)

Abstract: In order to reduce amount of data storage and improve processing capacity of the system, this paper proposes a new classification method of data source by combining phase synchronization model in network clustering with cloud model. Firstly, taking data source as a complex network, after the topography of network is obtained, the cloud model of each node data is determined by fuzzy analytic hierarchy process (AHP). Secondly, by calculating expectation, entropy and hyper entropy of the cloud model, comprehensive coupling strength is got and then it is regarded as the edge weight of topography. Finally, distribution curve is obtained by iterating the phase of each node by means of phase synchronization model. Thus classification of data source is completed. This method can not only provide convenience for storage, cleaning and compression of data, but also improve the efficiency of data analysis.

Key words: data classification; complex network; phase synchronization; cloud model

CLD number: TP274 **Document code**: A

Article ID: 1674-8042(2016)03-0266-06 **doi**: 10.3969/j.issn.1674-8042.2016.03.010

0 Introduction

As the time of big data is coming, a set of new challenges have sprung up, including data storage, data preprocessing and data fusion[1-2]. The classification of data source can effectively reduce amount of data storage, which is the precondition of data preprocessing and fusion decision. Due to diversity of data types (i.e. text data, numerical data and image data) and description methods in big data environment, undoubtedly, it increases difficulty of data processing and causes waste of storage space, therefore it is valuable to make a study of classification of data source for big data.

In fact, big data can be regarded as a complex network composed of various data sources, in which data source nodes are connected in a specific way. Therefore, the big data can be abstractly constructed into a topology structure chart, in which nodes represents data sources and edges represent correlation between data sources. At present, there have been a ton of research on complex network, such as community detection, image segmentation, and so on[3-4]. Most of them used the dynamics characteristics of complex network to find the relationship between seemingly unrelated nodes and then classify them. In addition, for various data sources of big data, each data source is a "big data" that sometimes is qualitative and of uncertainty, therefore we need to find an effective method to describe the data sources uniformly for the purpose of an effective classification. Cloud model not only can realize convertion of uncertainty information from qualitative to quantitative[5-6], but also can use membership to represent its credibility, which provides a convenience for the measurement of uncertainty information.

Received date: 2016-06-06

Foundation items: National Natural Science Foundation of China (No. 61171057, No. 61503345); Science Foundation for North University of China (No. 110246); Specialized Research Fund for Doctoral Program of Higher Education of China (No. 20121420110004); International Office of Shanxi Province Education Department of China, and Basic Research Project in Shanxi Province (Young Foundation)

Corresponding author: WANG Xiao-xia (wangxiaoxia@nuc.edu.cn)

To sum up, making use of the dynamic characteristics of complex network, this paper proposes a new classification method of data sources with the combination of phase synchronization model for network clustering and cloud model. The main idea of this method is that the distribution of each node data source is transformed into the problem of cloud model. By means of iteration of phase synchronization model, all nodes are classified.

1 Phase synchronization model of big data network

1.1 Dynamics characteristics of big data network

Big data includes the data with different types, different characteristics and different description methods and big data network is actually a complex network system composed of the data nodes with complementarity and similarity. Generally, the data of the network and the relationship between data information are bidirectional. Furthermore, for convenience the bidirectional relationship between the data nodes of the network is regarded as the same in this paper. If nodes represent different data sources in big data, an edge between two nodes is considered that there exists relationship between them and they are adjacent to each other. Edge weights are mainly used to describe the degree of similarity between two data source nodes. The greater its value, the greater the similarity between two nodes and the closer the relationship between two nodes. Therefore, this network can also be named as weighted complex networks of big data.

The dynamic characteristics of big data network means under the conditions influenced by outside world or triggered by data source nodes, the data information of each node changes or conection matrix between nodes varies, which causes a qualitative change of the entire network under some constraint conditions or randomly.

1.2 Network phase synchronization model

The big data network can be descried by undirected graph $G(\boldsymbol{N},\boldsymbol{E})$, among them, $\boldsymbol{N}$ is the set of nodes, $|\boldsymbol{N}|$ is the number of nodes, and $\boldsymbol{E}$ is the set of edges. Kuramoto model is the most representative method for phase synchronization[7-8]. It relates synchronous dynamic behavior and non-equilibrium phase change and can be expressed as

$$\frac{\mathrm{d}\theta_i}{\mathrm{d}t} = w_i + \frac{K}{|N|}\sum_{j=1}^{|N|} e_{ij}\sin(\theta_j - \theta_i), \quad i = 1,2,\cdots,|N|, \tag{1}$$

where θ_i is the phase of the i-th node; w_i is the inherent frequency of the i-th node; K is the coupling strength; $e_{ij} \in \{0,1\}$ represents the relationship between nodes i and j, if there is a edge between two nodes, $e_{ij}=e_{ji}=1$, otherwise $e_{ij}=e_{ji}=0$; For the diagonal node, $e_{ii}=0$.

After differential treatment of Eq. (1), the phase synchronization model based on discrete time iteration is got as

$$\theta_i(k+1) = \theta_i(k) + \frac{K}{|N|}\sum_{j=1}^{|N|} e_{ij}\sin(\theta_j(k) - \theta_i(k)), \tag{2}$$

$$K = \begin{cases} K_{\mathrm{p}} > 0, & e_{ij} \in \boldsymbol{E}, \\ K_{\mathrm{n}} < 0, & e_{ij} \in \bar{\boldsymbol{E}}, \end{cases} \tag{3}$$

where $i=1,2,\cdots,|N|$; $\theta_i(k)$ is the phase of the i-th node after k-th times interations; K is the coupling strength; $\boldsymbol{E}=[e_{ij}]$ is adjacency matrix; K_{p} is the coupling strength with edge between nodes i and j, also known as positive coupling strength; K_{n} is the coupling strength without edge between nodes i and j, also known as negative coupling strength.

It should be noted that K_{p} makes the phases of data source nodes with similar characteristics close to each other, especially, the greater the value, the faster synchronization; K_{n} makes the phases of data source nodes with different features away from each other, the greater the value, the faster the phases are away from each other.

The initial phases of $|V|$ nodes are randomly generated between $[0,2\pi]$, and then they are substituted into Eq. (2) for phase iteration. Because there is coupling relationship between data source nodes, the phases of data source nodes with big similarity tend to interconsistency, but the phases of data source nodes with little or without similarity will be away

from each other.

2 Classification of data sources in big data network

2.1 Description of data source

In order to find the coupling relationship of the nodes, we need to make a unified description for data source information. In this paper, cloud model is used to describe the uncertainty information and the definition of cloud model is as follows.

Definition 1 (cloud model [9]): Supposing that a random implementation of a qualitative concept C in the domain of discourse U is $x \in U$, if the uncertainty of x to C is $\mu(x) \in [0,1]$, where $\mu(x)$ is a random number with certain stability, the distribution of x in the domain of discourse U is called the cloud, marked as $C(X)$, where x is called cloud droplet.

Cloud model generally uses expectation, entropy and hyper entropy to describe the uncertainty of C. To get the cloud model, the fuzzy analytic hierarchy process (AHP) can be used[1].

2.2 Data classification algorithm based on network phase synchronization and cloud model

The basic principle of data classification based on network phase synchronization and cloud model is described as follows: Firstly, fuzzy AHP is used to obtain the cloud model of each data block and then the expectation, entropy and hyper entropy are calculated for the determination of the coupling strength between nodes[10]; Afterwards, the adjacency matrix of the network under test is produced based on the coupling strength, and the initialization phase is given; Finally, using the iterative phase synchronization model, the phase distribution curve is drawn and the data sources are classified. The detailed process is described as follows.

1) Use fuzzy AHP to obtain the cloud model of N data blocks. To facilitate calculation, assuming that cloud model obeys the triangular distribution, that is $[a-\alpha, a, a+\beta]$, then expectation M, entropy S and hyper entropy H of cloud model are calculated by

$$M = \int_0^1 [a - a(1-\lambda) + a + \beta(1-\lambda)]\lambda \mathrm{d}\lambda, \tag{4}$$

$$S = \frac{\pi}{2}\int_0^1 \{ | a - a(1-\lambda) - M | + | a + \beta(1-\lambda) - M | \lambda \mathrm{d}\lambda, \tag{5}$$

$$H = \sqrt{\left| \int_0^1 \{[a - a(1-\lambda) - M^2]^2 + [a + \beta(1-\lambda) - M]^2\}\lambda \mathrm{d}\lambda - M^2 \right|}. \tag{6}$$

2) To build a network, using N data blocks as network nodes, the coupling strength between nodes is calculated by

$$K_{e_{ij}} = \sqrt[3]{(M_j - M_i)(S_j - S_i)(H_j - H_i)}, \tag{7}$$

where i and j are two different nodes among the nodes to which N data blocks correspond.

3) According to the coupling strength between nodes, the average coupling strength of the whole network nodes can be got by

$$\overline{K} = \sum_{\substack{i,j \in [0,N] \\ i \neq j}} K_{e_{ij}}. \tag{8}$$

If $K_{e_{ij}} \geqslant \overline{K}$, there exists the edge e_{ij} between node i and node j, otherwise e_{ij} does not exist. If the edge exists between two nodes, the corresponding element in the adjacency matrix is 1; otherwise it is 0. The adjacency matrix $\boldsymbol{E} = [e_{ij}]_{N \times N}$ is obtained.

4) For big data network, whose node number is N, first of all, set K_p and K_n according to the actual situation, randomly generate N numbers in $[0, 2\pi]$ as the initial phases $\theta_1, \theta_2, \cdots, \theta_N$ of network nodes; then according to iterative phase synchronization model, when $\theta_i(k) - \theta_i(k-3) \leqslant 0.01$, the iteration stops, where $\theta_i(k)$ represents iteration phase of node i after the k-th iteration and total number of interation is m.

5) Divide $[0, 2\pi]$ into n_1 sub-intervals averagely, count up the number of nodes in each interval and draw phase distribution curve using n_1 as horizontal ordinate and the number of the nodes which fall in the intervals as vertical ordinate, finally the phase curve slope of the i-th ($i = 1, 2, \cdots, N$) node after sta-

bilization is calculated by

$$s_i = \frac{\theta_i(m) - \theta_i(m-3)}{3}. \tag{9}$$

Taking the number of wave trough as the number of classification, all the nodes between adjacent troughs on the curve are counted and classied as one class. Hence, n nodes are divided into L classes 1 and L is equal to the number of troughs.

6) For each of L classes, the following operations are carried out: Firstly, $[-0.01, 0.01]$ is divided into n_2 sub-intervals and the number of nodes whose slope falls in the sub-interval is counted. Secondly, taking n_2 as horizontal ordinate and the number of nodes whose slope falls in the subinterval as vertical ordinate, the distribution curve of the slope is drawn. Afterwards, the nodes between adjacent troughs on the curve are counted and classified as one class. Finally, the number of data source classification is equal to the total number of the troughs on the curve. As a result, all the nodes are divided into L' classes.

3 Application

In order to verify the feasibility and effectiveness of the proposed method, we take traffic data from a region of Shanxi Province as an example in this paper. The data include more than 15 000 records produced by more than 8 300 taxis from a taxi software, more than 300,000 taxi companies operating data, 1.3 million bus card charge data, 230 000 bus traveling recorder data and operating data, daily public bicycles platform credit card data and the record of population who walk healthily by app "Codoon". By analyzing a large amount of data, the proportion by taxi, bus, bicycle and walking on foot is 40.57 : 30.67 : 12.19 : 16.57.

This paper classifies and analyzes the above data to verify the proportion of population by taxi, bus, bicycle and walking on foot in Taiyuan city.

The data above are represented by N_i, $i=1,2,3,4,5,6,7$, among them, N_1, N_2 and N_3 are taxi data, N_4 and N_5 are bus data, N_6 is a public bicycle platform credit card data, G is walking people data by "Codoon" software. The proposed method is used to classify the above seven classes of data to verify the effectiveness of the method.

1) Work out the property values of all data blocks and obtain the cloud model of all data sources by means of the fuzzy AHP. This paper takes the triangular cloud distribution as an example to introduce the process, and the expectation M, entropy S and hyper entropy H of $N_i(i=1,2,3,4,5,6,7)$ are calculated by

$$M = [M_1, M_2, M_3, M_4, M_5, M_6, M_7] = [0.557\,2, 0.467\,7, 0.443\,3, 0.207\,0, 0.345\,3, 0.137\,9, 0.077\,4],$$

$$E = [E_1, E_2, E_3, E_4, E_5, E_6, E_7] = [0.400\,2, 0.355\,1, 0.374\,8, 0.190\,7, 0.196\,7, 0.099\,8, 0.044\,5],$$

$$H = [H_1, H_2, H_3, H_4, H_5, H_6, H_8] = [0.188\,5, 0.157\,5, 0.154\,7, 0.099\,5, 0.070\,0, 0.014\,9, 0.007\,0].$$

2) Taking these seven data blocks as seven network nodes, Eq. (7) is used to determine the coupling strength $K_{e_{ij}}$ between the nodes.

Taking the coupling strength between the network nodes as weight, the coupling strength matrix can be got as

$$[K_{e_{ij}}] = \begin{bmatrix} 1 & 0.050\,0 & 0.046\,1 & 0.186\,9 & 0.172\,2 & 0.144\,5 & 0.117\,9 \\ -0.050\,0 & 1 & -0.011\,0 & 0.135\,5 & 0.119\,3 & 0.100\,7 & 0.087\,2 \\ -0.046\,1 & 0.011\,0 & 1 & 0.133\,9 & 0.113\,9 & 0.098\,7 & 0.069\,8 \\ -0.186\,9 & -0.135\,5 & -0.133\,9 & 1 & 0.029\,0 & 0.024\,0 & 0.015\,9 \\ -0.172\,2 & -0.119\,3 & -0.113\,9 & -0.029\,0 & 1 & 0.009\,8 & 0.007\,5 \\ -0.144\,5 & -0.100\,7 & -0.098\,7 & -0.024\,0 & -0.009\,8 & 1 & 0.005\,6 \\ -0.117\,9 & -0.087\,2 & -0.069\,8 & -0.015\,9 & -0.007\,5 & -0.005\,6 & 1 \end{bmatrix}.$$

3) There are seven numbers generated randomly in $[0, 2\pi]$, and they are taken as the initial phases of

network nodes. Supposing that the positive feedback parameter K_p is 3 and negative feedback parameter K_n is -1, as well as Eq. (2) is used for phase iteration of five network nodes, when all nodes satisfy $\theta_i(k)-\theta_i(k-3)\leqslant 0.01$, the total number of iterations is $m=34$. Taking the 34th generation nodes phase, dividing $[0, 2\pi]$ into 50 sub-intervals and counting the number of node phases within every sub-intervals, the label-nodes distribution phase diagram is drawn, as shown in Fig. 1. It can be seen that two types of data (taxis and buses) are consolidated into a set of data due to the overlap of phase.

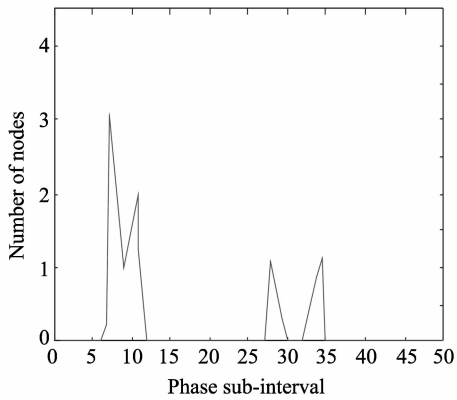

Fig. 1 Phase distribution diagram

In order to calculate the proportion of Taiyuan people by taxis and by buses everyday, the influence degree of slope on node is used for the division of slope sub-intervals and then for data classification. n_2 has linear relationship with it. Proper number of slope sub-intervals is beneficial for better data classification. Eq. (5) is used to calculatce the slopes of the m-th phase and the $(m-3)$-th phase of all the nodes, finally five slope values are got.

The interval $[-0.01, 0.01]$ is divided into four sub-intervals, the number of nodes whose slopes fall into every sub-interval. Taking n_2 as horizontal ordinate and the number of the nodes that fall into sub-interval as vertical ordinate, the slope distribution curve is drawn. All the nodes between adjacent troughs on each curve are counted and classified as the same class, as shown in Fig. 2.

The four types of data (taxis, buses, bicycles and on foot) are well classified, thus, according to the results, the proportion of people living in Taiyuan who travel by taxi, bus, bicycle and on foot is 3 : 2 : 1 : 1. The results are the same as the original statistic approximation, moreover the efficiency of the data analysis has been improved.

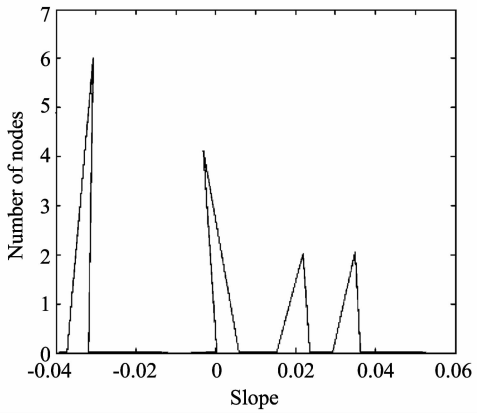

Fig. 2 Slope distribution of data nodes

4 Conclusion

In this paper, a data classification method is proposed based on network dynamics analysis and cloud model, which makes big data storage, cleaning and compression more convenient. The method is validated with an example of the number of people by taxi and by bus every day in Taiyuan city. The results show that the proposed data classification method can classify the data sources effectively.

References

[1] LI De-ren, ZHANG Liang-pei, XIA Gui-song. Automatic analysis and mining of remote sensing big data. Acta Geodaetica Et Cartographica Sinica, 2014, 43(12): 1211-1216.

[2] HU Guang-yong. Research of data security storage strategy based on cloud computing. Computer Measurement & Control, 2011, 19(10): 2539-2541.

[3] DU Yue, DU Wen, CHEN Shuang. Application of complex network theory to urban transportation network analysis. Urban Transport of China. 2009, 7(1): 57-65.

[4] Lee K E, Lee J W, Hong B H. Complex networks in a stock market. Computer Physics Communications, 2007, 177(1/2): 110-111.

[5] LI De-yi. Uncertainty artificial intelligence. Beijing: National Defense of Industry Press, 2005: 171-177.

[6] XU Ze-shui. Fuzzy aggregation operators. IEEE Transactions on Fuzzy Systems, 2007, 15(6): 1179-1187.

[7] WU Jian-she, JIAO Li-cheng, JIN Chao. Overlapping community detection via network dynamics. Physical Re-

view E, 2012, 85(2): 0161151-0161157.

[8] JIAO Yang. Phase synchronization model and its applications. Master thesis. Xi'an: Xidian University, 2014.

[9] CHEN Gui-lin. Uncertain model of qualitative/quantitative information transformation-cloud model. Application Research of Computers, 2010, 27(6): 2006-2010.

[10] LI Hong-xian, Mohamed A H, ZHEN Lei. Risk identification and assessment of modular construction utilizing fuzzy analytic hierarchy process and simulation. Canadian Journal of Civil Engineering, 2013, 40(12): 1184-1195.

基于网络动力学分析和云模型的数据分类方法

王肖霞，杨风暴，梁若飞，张文华

（中北大学 信息与通信工程学院，山西 太原 030051）

摘　要： 为了降低数据的存储量和提高系统的处理能力，本文将网络聚类中的相位同步模型和云模型相结合，提出了一种新的数据源的分类方法。首先，将数据源看作一个复杂网络，获得其拓扑结构图，利用模糊层次分析法确定各节点数据源的云模型；然后，计算云模型的期望值、熵和超熵，获得综合耦合强度，将其作为边权；最后，利用相位同步模型对各节点相位进行迭代，获得其分布曲线，进而对各数据源进行分类。本文所提出的方法不仅为数据的存储、清理和压缩提供方便，而且提高了数据的分析效率。

关键词： 数据分类；复杂网络；相位同步；云模型

引用格式： WANG Xiao-xia, YANG Feng-bao, LIANG Ruo-fei, et al. Data classification method based on network dynamics analysis and cloud model. Journal of Measurement Science and Instrumentation, 2016, 7(3): 266-271. [doi: 10.3969/j.issn.1674-8042.2016.03.010]